Part I

Theory

Chapter 1

The Nature of Meaning

If we want to extract meaning automatically from text, we first need to investigate what 'meaning' actually is. What does it mean for a word to mean something? How is a particular word able to convey a particular content? And how is this content built up? What is, in short, the nature of meaning?

1.1 Theories of meaning

In the course of history, the nature of meaning has been one of the major issues in the philosophical debate. The issue was first raised in the ancient Greek world, and was subsequently tackled by numerous philosophers. In the 19th century, meaning also entered the realm of linguistics – first in the context of diachronic linguistics,[1] later also as a synchronic study. In the following paragraphs, we briefly discuss the different theories of meaning (and their relation to reality) that have been proposed both in philosophy and linguistics, and assess their potential to serve within computationally implemented procedures of meaning extraction.

1.1.1 Referential theory of meaning

In a referential theory of meaning, the meaning of a particular word is regarded as a pointer to the designated object in the real world. The meaning of a word

[1] Christian Karl Reisig proposed the study of 'Semasiologie' in 1825, Michel Bréal coined the term 'sémantique' in 1883.

is what it refers to. If we utter a word like *apple*, we refer to an actual apple (or the set of all apples) in reality. Intuitively, a referential theory of meaning seems very appealing. If parents want to teach their children the meaning of a word like *apple*, chances are pretty high that they will point to an actual apple – or a picture of one. At first sight words indeed seem no more than references to things (entities, actions or relations) existing in the outside world. There are, however, a number of problems with such a referential theory of meaning. The theory is able to account for what is generally called the *denotation* or *extension* of words, but fails to describe other semantic characteristics, generally referred to as *connotation* or *intension*. The German philosopher Gottlob Frege (1848–1925) illustrated this deficiency with a by now well-known example. Compare the following sentences:

(1) The morning star is the morning star.

(2) The morning star is the evening star.

Both *morning star* and *evening star* refer to the same entity, viz. the planet Venus, which might be visible either in the morning or in the evening (depending on the relative position of Venus and the earth). Sentences (1) and (2), however, significantly differ in meaning. Sentence (1) expresses a simple tautology, whereas sentence (2) expresses a new and important astronomical truth. Sentences (1) and (2) do not mean the same thing, but a referential theory of meaning does not account for the difference between them.

Frege's solution to the morning/evening star paradox was to make a distinction between *Sinn* (sense) and *Bedeutung* (reference). *Bedeutung* is the object that the word refers to, whereas *Sinn* is the cognitive representation of the object. By making this distinction, it is possible for words to have a different sense but the same referent (as in the paradox above).

The referential theory of meaning has been popular with logicians (e.g. the young Wittgenstein and Bertrand Russell). It provides a parsimonious and straightforward model of meaning, but the previous examples have shown that it is incapable of capturing all aspects of meaning. Moreover, it is unclear how we ought to proceed in order to extract these 'meaning references' in a computational way. The theoretical problems as well as the practical drawbacks make the referential theory rather unattractive for the computational extraction of meaning.

1.1.2 Mentalist theory of meaning

Another solution – one that has been very popular throughout the history of philosophy, starting with the Greek philosopher Plato – is to represent meaning

exclusively as ideas. A mentalist theory of meaning associates the meaning of a particular word with a particular idea in the human mind. This theory effectively solves the morning/evening star paradox: The morning star might be the same thing as the evening star in reality, but the *idea* of the morning star and the evening star may very well differ. The question that immediately follows is what this notion of *idea* actually entails. Surely, it cannot be the mental representations that are present in each individual person. These mental representations differ a lot among different persons. If one person hears the word *strawberry*, an image of an appetizing dessert plate – possibly covered with lots of whipped cream – might pop up. Another person might prefer them with powder sugar, and another one without any topping at all. Or one might even be disgusted by the idea of strawberries, because of a severe allergic reaction in the past. To be practically usable, the *ideas* need to have some generality, exceeding the individual level. But it is difficult to achieve this generalization without resorting to the notion of idea in the platonic sense, that is somehow mysteriously present in people's minds. This is not the direction we want to venture into, especially if we want to implement semantics in a computational way. If we want a sound theory of semantics that can be implemented computationally, we will need a theory that is not dependent on reference or ideas.

1.1.3 Behavioural theory of meaning

The vagueness and non-generality that inevitably seems to surround the mentalist view has led people to abandon the mentalist theory of meaning in favour of a theory that sticks to 'observable' facts. Inspired by the behaviourist movement that became popular within the field of psychology, the American linguist Leonard Bloomfield defines meaning of a linguistic form as 'the situation in which the speaker utters it and the response which it calls forth in the hearer'. (Bloomfield, 1933, p. 139). The meaning of a word is thus reduced to the speaker's stimulus that elicits its use, and/or the hearer's response to that word.

Although the behavioural theory of meaning claims to overcome the vagueness of ideas in the mentalist view, it seems almost as problematic as the theory it opposes. There is a plethora of different stimuli that elicit the same word, and the number of different responses evoked by that word is equally high. Take, for example, a word like *jazz*. In some situations, a person might utter the word to indicate they would like to hear some jazz tunes. In other situations, they might utter the word to approve – or disapprove – of the music they are listening to at that moment. And one odd person – not particularly familiar with different music styles – might even utter *jazz* when in fact they are listening to hip hop. Similarly, people's

reactions to the word *jazz* may differ quite a lot. One person might turn on the radio and look for a suitable radio station, another one might start nodding their head whistling a Duke Ellington tune, while yet another might make an unhappy face and stick out their tongue. Every language utterance has a similar abundance of stimuli eliciting it, and a similar abundance of responses following it. This makes it practically impossible to describe the meaning of a particular word in terms of the utterance's stimuli and responses.

Moreover, behaviourists have a rather vague and untenable view of what this behaviourist meaning description practically should look like. In his main textbook on linguistics, *Language*, Bloomfield notes:

> The situations which prompt people to utter speech, include every object and happening in their universe. In order to give a scientifically accurate definition of meaning for every form of a language, we should have to have a scientifically accurate knowledge of everything in the speaker's world. (Bloomfield, 1933, p. 139)

Bloomfield himself acknowledges that

> ... the statement of meanings is therefore the weak point in [behavioural] language-study, and will remain so until human knowledge advances very far beyond its present state. (Bloomfield, 1933, p. 140)

Bloomfield also deems it necessary to 'resort to makeshift theories' whenever scientific description is impossible – one of those theories being a referential theory of meaning.

In addition to the theoretical and practical drawbacks associated with a behaviourist's description of meaning, the theory obviously doesn't stand a chance to function within a computational framework. In order to implement meaning in a behavourial, computational framework, a computer should be able to observe, interpret and classify human stimuli and responses. The current state of artificial intelligence does not allow such complex cognitive computations just yet.

1.1.4 Use theory of meaning

A radically different theory of meaning qualifies the meaning of an expression as its use in a language system. A use theory of meaning does not refer to an external entity (a referent, an idea, or stimuli and responses) to qualify a word's meaning, but instead qualifies the meaning of a word as the value it gets through

the (linguistic) system in which it is used. It was Wittgenstein who famously noted that 'the meaning of a word is its use in the language' (Wittgenstein, 1953).

The use theory of meaning differs radically from the previous theories of meaning. In the previous theories, there is an existing order of things (a 'meaning') outside of the language system; the words of a language are used to talk about existing entities, but entities and words belong to two different classes, and there is no influence between the two classes. A use theory of meaning, on the other hand, advances a system in which meaning is defined and constructed within the language itself.

The first person to explore this radically different view in the context of linguistic theory was the Swiss linguist Ferdinand de Saussure (1857–1913), who is the founding father of the linguistic movement nowadays known as structuralism. In his most important work, the *Cours de Linguistique Générale (Course in General Linguistics)*, Saussure lays out the foundations for a differential view on language. Saussure defines a linguistic sign as a combination of the *signifiant* ('signifier') – representing the sound form of the sign – and the *signifié* ('signified') – representing the linguistic meaning of the sign.

According to Saussure,

> . . . la langue est un système dont tous les termes sont solidaires et où la valeur de l'un ne résulte que de la présence simultanée des autres . . . [language is a system in which all terms are equal, and in which the value of one is only the result of the simultaneous presence of the others] (Saussure, 1916)

This quote represents Saussure's structuralist view on language: the linguistic meaning of a sign is not a given, existing truth in the outside world, but it is defined in terms of its use in particular contexts (and its non-use in other contexts). The meaning of a particular word is not an independent or transcendental fact, but it is defined within a network of different embedded meanings, which in turn get their values from their position in the network of meanings.

The structuralist view on language has been further developed by a number of linguists, one of the most notable being Zellig Harris. Harris advocated a distributional method for linguistic research: linguistic elements (words, but also morphemes or phonemes) can be investigated by looking at the way they are distributed in language. As such, the distributional method is also able to discover the semantic properties of a word. Harris notes:

> The fact that, for example, not every adjective occurs with every noun can be used as a measure of meaning difference. For it is not merely

> that different members of the one class have different selections of members of the other class with which they are actually found. More than that: if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution. (Harris, 1954, p. 156)

The hypothesis that semantically similar words tend to occur in similar contexts has been coined the DISTRIBUTIONAL HYPOTHESIS in subsequent work, and Harris' work is often cited as the main source of inspiration. The distributional hypothesis not only turns out to provide a sound basis for meaning description, it also provides a suitable starting point for an implementation in a computational framework, as will be shown in chapter 2.

1.2 Context

In the first part of this chapter, we have shown that a use theory of meaning is a sound basis for the computational extraction of lexico-semantic meaning: the sum of a word's contexts is a good indicator of the word's use, and hence its 'meaning'. In the second part of this chapter, we focus on the notion of context. The context of a particular word can be interpreted in a number of ways: the context might be the document the word appears in, it might be a window of words around a particular word, or it might be the syntactic context in which the word takes part. In this section, we will have a look at these different kinds of context, investigate which parameters are involved, and examine how different contexts might be useful for lexico-semantic knowledge extraction. The next chapter will then investigate how these various contexts can be formalized and implemented in a computational way.

1.2.1 Document-based context

First of all, a particular word always appears in a particular document. This gives rise to our first instantiation of the distributional hypothesis:

Hypothesis 1. *Words are semantically similar if they appear in similar documents.*

Words that appear in the same documents tend to be thematically related: texts usually focus on one particular topic (or a few topics), so that the majority of content words is related to these topics. Take for example the three newspaper

paragraphs in figure 1.1, taken from the MEDIARGUS newspaper corpus, a 1.4 billion word corpus of (Belgian) Dutch newspaper texts.

The three paragraphs all contain words related to the medical domain (printed in boldface). Note that a word like *patiënt* 'patient' appears in all three documents. Likewise, words like *dokter* 'doctor' and *arts* 'doctor' appear in the same documents. In a similar vein, words related to another topic – say economics, soccer, or rock music – appear together in the same documents. If such related words appear in the same documents sufficiently frequently, a computer algorithm might be able to infer that they are indeed semantically related.

The main parameter to be set is the size of the document context. This will depend on the corpus used and the application in mind. In a newspaper corpus, the unit might be an article, or a paragraph. When using a web corpus, one might consider a particular web page as document context.

1.2.2 Window-based context

Secondly, a particular word appears within the context of other words in its vicinity, which brings us to our second instantiation of the distributional hypothesis:

Hypothesis 2. *Words are semantically similar if they appear within similar context windows.*

Below are some examples taken from the Twente Nieuws Corpus (TWNC), a 500M word corpus of Dutch newspaper texts. Examples (3) to (5) all contain the word *courgette* 'zucchini'. Examples (6) to (8) all contain the word *aubergine* 'eggplant'.

(3) Kies eens voor tomaat, paprika, dun geschaafde **courgette** en plakjes
choose once for tomato pepper thin sliced zucchini and slices
rauwe champignons.
raw mushrooms
Pick a tomato, pepper, thinly sliced zucchini and slices of raw mushrooms for once.

(4) Serveer met pasta en gebakken groente, zoals paprika, **courgette** en
serve with pasta and fried vegetable like pepper zucchini and
tomaat.
tomato
Serve with pasta and fried vegetables, such as pepper, zucchini and tomato.

Uit het onderzoek blijkt ook dat slechts de helft van de **patiënten** naar de **dokter** gaat. Veertig procent praat over z'n probleem met vrienden. Maar 20 procent van de **patiënten** heeft het er met niemand over. **Patiënten** die hun **kwaal** voor de buitenwereld verbergen zeggen 'de juiste woorden niet te vinden om hun toestand te beschrijven', of 'zich te schamen over hun toestand'. Sommigen hadden zelfs schrik om er met iemand over te praten. Volgens het onderzoek dient de reden waarom mensen er niet over praten ook bij de **arts** te worden gezocht.

The research also shows that only half of the patients goes to the doctor. Forty percent talks about their problem with friends. But twenty percent of the patients does not talk to anybody. Patients hiding their condition for the outside world claim 'not to find the right words to describe their situation', or 'to be ashamed of their situation'. Some were even frightened to talk to someone about it. The research shows that the reason why people are not talking about it also has to be sought with the doctor.

In heel wat gevallen kunnen **dokters** zich beter concentreren op de oorzaken van de **pijn** in plaats van op de **behandeling**, zo wil de nieuwe denktrant in de **medische** wereld. **Operaties** zijn uitzonderlijk, het gros van de **patiënten** is gebaat bij de zogenaamde conservatieve (**niet-operatieve**) **therapieën**. De topper is **oefentherapie** onder begeleiding van een **kinesitherapeut**.

In many cases, doctors would better concentrate on the causes of the pain instead of the treatment, that is the new way of thinking in the medical world. Surgeries are exceptional, the majority of the patients benefits from so-called conservative (non-surgical) therapies. The top therapy is training therapy, coached by a physiotherapist.

Zaterdagvoormiddag werd ingebroken in de wagen van een **arts** die op huisbezoek was bij een **patiënt**. De dader sloeg een ruit van de wagen stuk, vond de **doktersjas** en nam een aantal **spuiten** mee.

Saturday morning, the car of a doctor visiting a patient was burgled. The offender broke a window, found the doctor's coat and took a couple of injections.

Figure 1.1: Three document paragraphs from different newspapers – all containing words from the medical domain – extracted from the MEDIARGUS corpus

(5) Deze Indiase currysoep (mulligatawny) krijgt een zomers tintje
this Indian curry soup mulligatawny gets a summery touch
door de **courgette** en paprika.
through the zucchini and pepper
This Indian curry soup (mulligatawny) gets a summery touch because of the zucchini and pepper.

(6) Snijd groenten, zoals paprika, **aubergine** en ui in kleine stukjes.
cut vegetables like pepper eggplant and onion in small pieces
Cut vegetables, such as pepper, eggplant and onion in small pieces.

(7) Natuurlijk, in Purmerend verkopen ze ook tomaten, en **aubergine**,
of course in Purmerend sell they also tomatoes and eggplant
en paprika.
and pepper
Of course, tomatoes, eggplants and peppers are also sold in Purmerend.

(8) Voeg **aubergine**, aardappelen, bataat (zoete aardappel) en paprika
add$_{verb}$ eggplant potatoes bataat sweet potato and pepper
toe.
add$_{particle}$
Add eggplant, potatoes, bataat (sweet potato) and peppers.

Note that *courgette* and *aubergine* in the examples above have a tendency to occur with the same words, such as *paprika* 'pepper', *tomaat* 'tomato', and *groente* 'vegetable'. Again, if such related words (like *courgette* and *aubergine*) occur with the same words (like *paprika* and *tomaat*) sufficiently frequently, a computer algorithm might be able to infer that they are indeed semantically related. Note that *courgette* and *aubergine* even do not have to occur together (although they might). It is their co-occurrence with other words that is indicative of their semantic relatedness.

A simple context window as described above is often called a bag of words context. This expression is used to indicate the fact that no order (or syntax) is taken into account; the ordered words are mixed together ('put together in one bag') so that their internal order is lost.

The main parameter to be set is the size of the window in which a word's context words occur. One might take into account a small window, in which only the left and right co-occurring word are used as context. A medium-sized window might use two or five words to the left and right of the word in question. A large window might take into account all context words that occur in the same sentence, or even in the same paragraph.

One can imagine that different context window sizes will lead to different kinds of semantic similarity. When using a small context window, an algorithm might be able to find tight semantic relationships: in a small context window, more closely related context words might appear in the word's vicinity, and the algorithm might even be able to discover some basic syntactic facts (e.g. the fact that a particular word appears with an article). When using a larger context window, more loosely related words might show up, and all order gets lost in the bag of words. Using larger windows, the algorithm might be more likely to discover topically similar words again. In the second part of this thesis, we will investigate whether this hypothesis is true.

1.2.3 Syntax-based context

Thirdly, a particular word always takes part in particular syntactic relations. This gives rise to our third and final instantiation of the distributional hypothesis:

Hypothesis 3. *Words are semantically similar if they appear in similar syntactic contexts.*

In this research, syntactic context will be instantiated in the form of dependency graphs; dependency graphs provide a theory-neutral instantiation of a sentence's syntax, since no particular grammatical framework is assumed. More specifically, this research will use dependency structures that conform to the guidelines for the *Corpus Spoken Dutch* (CGN, Hoekstra et al. (2001)). The dependency structures used in the CGN syntactic annotation have developed into a de facto standard for the computational analysis of Dutch (Bouma, van Noord, and Malouf, 2001) and they are used as output format of the Dutch dependency parser ALPINO (van Noord, 2006). Formally, a CGN dependency structure $D = \langle V, E \rangle$ is a labeled directed acyclic graph, with node labels V representing the categories (phrasal labels and POS labels) and edge labels E representing the dependency relations.

Figures 1.2 and 1.3 show dependency structures for two sentences from the MEDIARGUS corpus, parsed with ALPINO. Table 1.1 shows the set of dependencies that can be deduced from the structures. In our syntax-based models of semantic similarity (discussed in the next chapter), we will use these dependency triples as the input data.

Note that both *biertje*[2] and *wijn* appear as direct object of the verb *drink*. Again, if we look at a large number of sentences, we might notice that *biertje* and *wijn* appear in similar syntactic contexts.

[2] DIM indicates the word is a diminutive form.

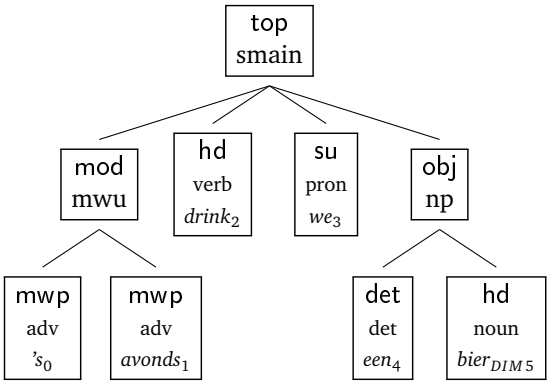

Figure 1.2: Dependency structure for the sentence *'s avonds drinken we een biertje* ('in the evening we'll drink a beer')

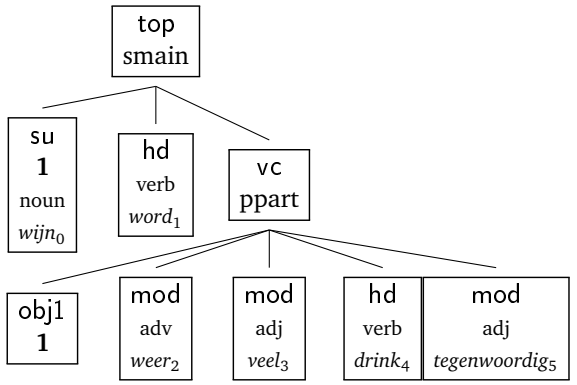

Figure 1.3: Dependency structure for the sentence *wijn wordt weer veel gedronken tegenwoordig* ('wine is drunk a lot again today')

One important parameter in syntax-based models is the set of dependency relations that will be incorporated into the model. A number of dependency relations that might be useful in distributional similarity models are given in table 1.2. These are the dependency relations that will be used in the syntax-based models presented in this thesis.

⟨drink	mod	's avonds	⟩
⟨drink	su	we	⟩
⟨drink	obj1	bier_{DIM}	⟩
⟨word	su	wijn	⟩
⟨drink	obj1	wijn	⟩
⟨drink	mod	weer	⟩
⟨drink	mod	veel	⟩
⟨drink	mod	tegenwoordig	⟩

Table 1.1: The set of dependency triples extracted from the two parses in figures 1.2 and 1.3

abbr.	**relation**	**example**
SU	subject	⟨*author*, SU, *write*⟩
OBJ1	direct object	⟨*wine*, OBJ1, *drink*⟩
OBJ2	indirect object	⟨*him*, OBJ2, *give*⟩
PC	prepositional complement	⟨*dog*, PC, *look_after*⟩
MOD	modifier	⟨*red*, MOD, *apple*⟩
PREDC	predicative complement	⟨*apple*, PREDC, *tasty*⟩
COO	coordination	⟨*apple*, COO, *pear*⟩
APP	apposition	⟨*London*, APP, *city*⟩

Table 1.2: Dependency relations used as contexts

1.3 Tight vs. topical similarity

We already briefly mentioned the difference between tight, synonym-like semantic similarity and more loosely related, topical similarity. With tight similarity, we indicate the fact that two words are very similar, i.e. there is a (near-)synonymous or (co-)hyponymous relationship between the two words. With topically similar words, we mean words that belong to the same semantic domain.[3]

The example below makes clear the difference between both kinds of similarity. Two sets of words are given that are semantically similar to the word *arts* 'doctor'. The first set contains words that are tightly similar to *arts*, containing synonyms (e.g. *dokter* 'doctor') and hyponyms (e.g. *chirurg* 'surgeon'). The second set of

[3] Often, the terms 'semantic similarity' and 'semantic relatedness' are also used to make a distinction between both kinds of similarity.

words is topically related to *arts*, containing words that all belong to the medical domain. The topically related words are related to the target word by more loose relationships, such as association and meronymy (part-whole relationships).

1. *dokter* 'doctor', *medicus* 'doctor', *huisarts* 'family doctor', *chirurg* 'surgeon', *specialist* 'specialist', *gynaecoloog* 'gynaecologist'

2. *patiënt* 'patient', *ziekte* 'disease', *diagnose* 'diagnosis, *behandeling* 'treatment, *ziekenhuis* 'hospital', *stethoscoop* 'stethoscope'

In the evaluation part (the second part of this thesis), we will not only try to evaluate the performance of the various models for the extraction of semantic similarity; we will also try to determine the nature of the similarity, i.e. whether the models are extracting tight, synonym-like similarity or more loosely related, topical similarity.